MATH 310 Project 2

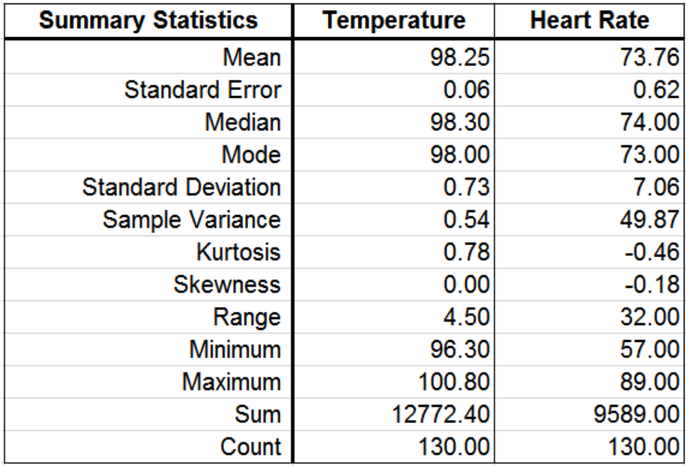
Kate Sanders

30 November 2017

1 Normal Temperature

I was given a data set with the body temperature, heart rate, and gender of $n = 130$ subjects, which mimicked the results of a 1992 article from the *Journal of the American Medical Association.* Using Excel's Data Analysis toolkit, I created a Summary Statistics Table that includes sample mean, variance, and standard deviation. These values can be seen in Table 1.

Table 1: Summary Statistics of Heart Rate and Body Temperature Data



Summary Statistics	Temperature	Heart Rate
Mean	98.25	73.76
Standard Error	0.06	0.62
Median	98.30	74.00
Mode	98.00	73.00
Standard Deviation	0.73	7.06
Sample Variance	0.54	49.87
Kurtosis	0.78	-0.46
Skewness	0.00	-0.18
Range	4.50	32.00
Minimum	96.30	57.00
Maximum	100.80	89.00
Sum	12772.40	9589.00
Count	130.00	130.00

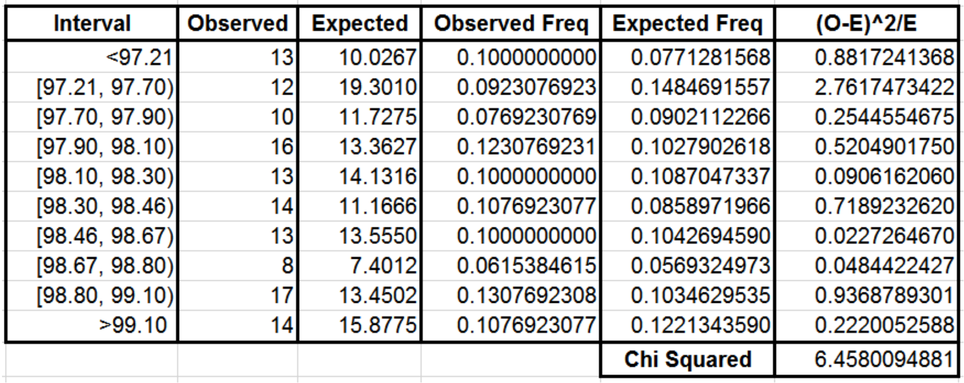
To see whether the distribution of temperature is approximately normal, I used the χ^2 test with $H_0 = N(98.25, 0.54)$. Observed and Expected values, as well as the χ^2 calculation can be seen in Table 2. These calculations were done in Excel. The sum of the sample χ^2 values was 6.458. Using Table IV of *Probability and Statistical Inference*, $\chi^2_{0.05}(9) = 16.92$. Since $6.458 < 16.92$, we fail to reject the null hypothesis, $H_0 = N(98.25, 0.54)$. Thus, the distribution of body temperatures is approximately normal.

A person's temperature would be considered "abnormal" if it lay outside of the 95% confidence interval of the mean. Since the distribution is approximately normal and n is large, the 95% confidence interval can be derived using the statistics from Table 1 as follows:

$$[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$
$$[98.25 - \Phi(z_{0.05/2})\frac{0.73}{\sqrt{130}}, 98.25 + \Phi(z_{0.5/2})\frac{0.73}{\sqrt{130}}]$$
$$[98.124, 98.375] \quad .$$

It is not surprising that range of the 95% confidence interval is small; this is a large sample size with a small variance.

Table 2: Grouped Body Temperature Data for χ^2 Calculation



Interval	Observed	Expected	Observed Freq	Expected Freq	(O-E)^2/E
<97.21	13	10.0267	0.1000000000	0.0771281568	0.8817241368
[97.21, 97.70)	12	19.3010	0.0923076923	0.1484691557	2.7617473422
[97.70, 97.90)	10	11.7275	0.0769230769	0.0902112266	0.2544554675
[97.90, 98.10)	16	13.3627	0.1230769231	0.1027902618	0.5204901750
[98.10, 98.30)	13	14.1316	0.1000000000	0.1087047337	0.0906162060
[98.30, 98.46)	14	11.1666	0.1076923077	0.0858971966	0.7189232620
[98.46, 98.67)	13	13.5550	0.1000000000	0.1042694590	0.0227264670
[98.67, 98.80)	8	7.4012	0.0615384615	0.0569324973	0.0484422427
[98.80, 99.10)	17	13.4502	0.1307692308	0.1034629535	0.9368789301
>99.10	14	15.8775	0.1076923077	0.1221343590	0.2220052588
				Chi Squared	6.4580094881

Since 98.6 Fahrenheit lies outside of the 95% confidence interval, it is not likely the true population mean for body temperature. To support this claim, I ran a 2-tailed test of the hypotheses $H_0 : \mu = 98.6$ and $H_1 : \mu \neq 98.6$. Since the distribution for body temperature is approximately normal and n is sufficiently large, the critical region can be found using:

$$|\bar{x} - \mu_0| \geq z_{\alpha/2}\frac{s}{\sqrt{n}}$$

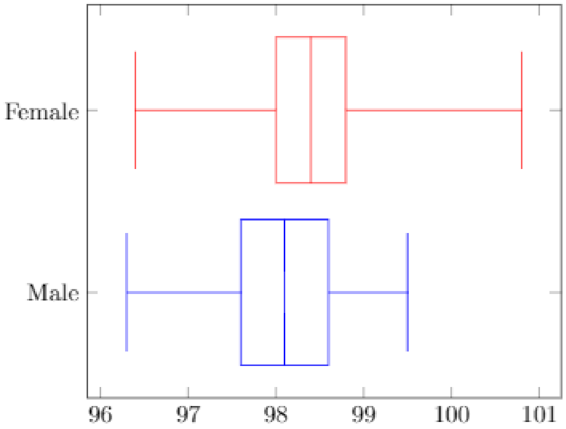
Using a significance value of $\alpha = 0.05$, $\Phi(z_{\alpha/2}) = 1.960$ according to Table Va of *Probability and Statistical Inference*. Using values from Table 1, I plugged in sample standard deviation, s; sample mean, $\bar{x}$; null hypothesis mean, μ_0; and number of samples, n. These values can be seen in Table 1.

$$|98.25 - 98.6| \geq 1.960\frac{0.73}{\sqrt{130}}$$
$$0.35 \geq 0.1254894614$$

Therefore, we reject the null hypothesis, $H_0 : \mu = 98.6$ degrees Fahrenheit.

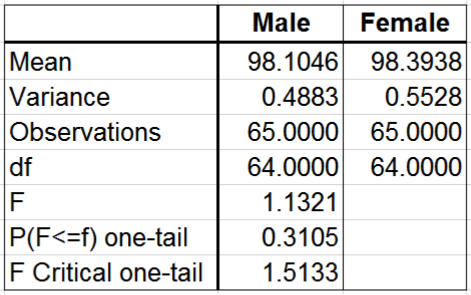
Looking at how body temperature differs with sex is interesting. A Box and Whisker plot of the male and female body temperature data can be seen in Figure 1. The median and range of the female data is larger than that of the male data. As seen in Table 4, the mean of both male and female body temperatures lay outside of the 95% confidence interval for body temperature, suggesting an underlying sex difference in the measurements.

Figure 1: Plot of Male and Female Body Temperatures in Fahrenheit



Though variance could be assumed equal since $n > 30$, I used the Two-Sample F-Test for Variance in Excel with the hypotheses $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$ and a significance level of $\alpha = 0.05$. The null hypothesis for the test was that the variances are equal. The results can be viewed in Table 3. Since $F < F_{Crit}$ and the P-value is sufficiently large, I fail to reject the null hypothesis

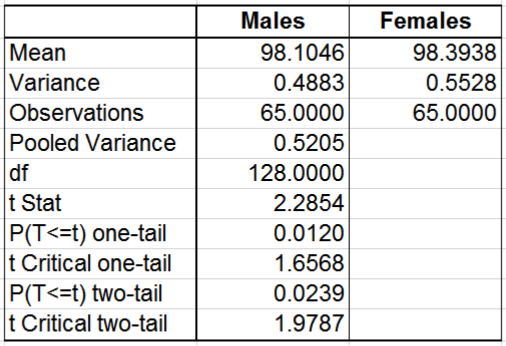
Table 3: Two-Sample F-Test for Variance



	Male	Female
Mean	98.1046	98.3938
Variance	0.4883	0.5528
Observations	65.0000	65.0000
df	64.0000	64.0000
F	1.1321	
P(F<=f) one-tail	0.3105	
F Critical one-tail	1.5133	

I then compared the means of male and female body temperature samples in Excel using the Two-Sample t-Test Assuming Equal Variances. Since the null hypothesis is that the means are equal, I used the two-tail values to determine whether or not I should reject the null hypothesis. The results of the test can be seen in Table 4. Since the two-tail p-value is less than the significance level $\alpha = 0.05$, I reject the null hypothesis. There is not sufficient evidence to support the claim that male and female patients have equal body temperatures.

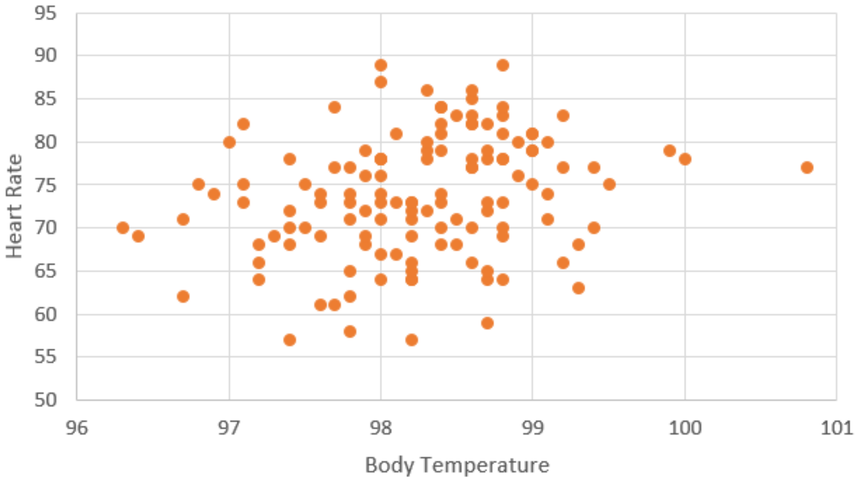
Table 4: Two-Sample t-Test Assuming Equal Variances



	Males	Females
Mean	98.1046	98.3938
Variance	0.4883	0.5528
Observations	65.0000	65.0000
Pooled Variance	0.5205	
df	128.0000	
t Stat	2.2854	
P(T<=t) one-tail	0.0120	
t Critical one-tail	1.6568	
P(T<=t) two-tail	0.0239	
t Critical two-tail	1.9787	

Lastly, I wanted to look for a relationship between body temperature and heart rate. A scatter plot of this data can be seen in Figure 2. To determine whether there is any statistically significant correlation in the data, I tested the hypothesis $H_0 : \rho = 0$ against $H : \rho \neq 0$ using a two-tailed test with a significance level $\alpha = 0.05$. Using Excel, I determined that the sample correlation coefficient is $r = 0.2536$. With over 100 degrees of freedom, $r_{0.025} = 0.1946$. This value was obtained from Table IX of *Probability and Statistical Inference*. Since $0.2536 > 0.1946$, I reject H_0. This means that the there is no significant relationship between a person's body temperature and heart rate.

Figure 2: Scatter Plot of Heart Rate and Body Temperature Data



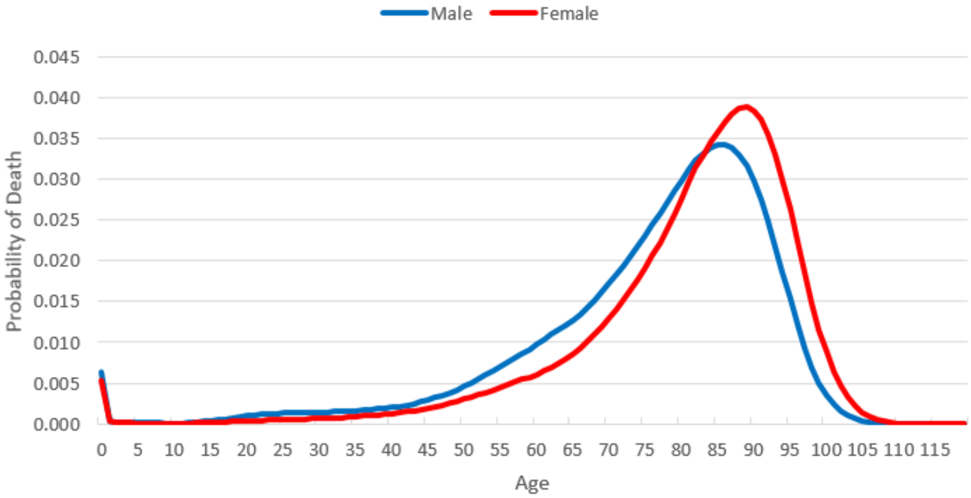
2 Mortality

The mortality rates of Americans are kept by the Social Security Administration. This data is split by sex and contains the conditional probability that an individual will die at age X. There is also a ”Number Lives” column for each gender showing how many people are alive at age X out of a hypothetical 100,000 person sample.

To construct a discrete probability distribution of this data, I used the ”Number Lives” column. To get the probability that a person of age X lived another year, I subtracted the ”Number Lives” value at age $(X+1)$ from ”Number Lives” value at age X and divided that value by 100,000. I checked the probability distribution for each sex by also creating a cumulative probability distribution column in Excel; as expected, this column's value reached 1 and leveled off. A line graph comparing the probability distributions of the mortality rate for men and women is shown in Figure 3.

There are several interesting characteristics evident in the graph of the distributions. Firstly, there is a sharp drop in mortality rates in the first few years after birth. Death rates for infants are high due to birth defects and a poor immune system. In the US, if a person lives past infancy, their probability of dying during childhood is quite low. The distributions are left skewed because people are more likely to die when they are older.

Figure 3: Probability of Death by Age for Males and Females

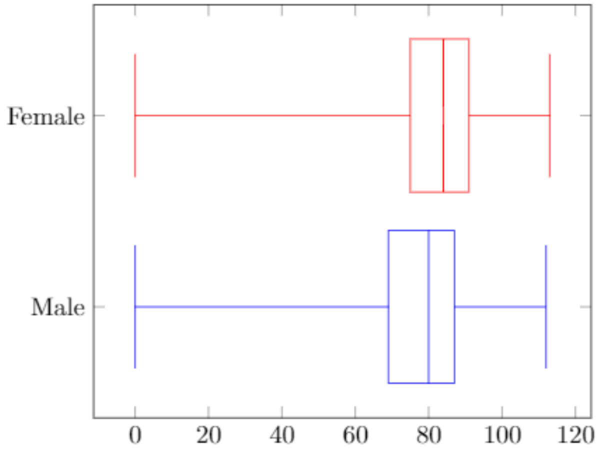


The is also a rise in the probability of death for males when they reach their twenties. At this time, men often engage in riskier behaviour. By age 20, there is a noticeable difference between the distributions of males and females. The maximum death probability is reached for men before women. This could be partially due to male's increased risk of heart disease. Though the maximum death rate for females occur later, the mode is larger. These statistics will now be examined in closer detail.

For men, the highest probability of death is 3.415% at 86 years of age; 68.642% of men die before the mode age. For women, the highest probability of death is 3.881% at 89 years of age; 66.469% of women die before the mode age.

Using the CDF and "Number Lives" column, I found the 5-number summary for both males and females. For both men an women, $Q_0 = 0$. For females, $Q_1 = 75, Q_2 = 84, Q_3 = 91, Q_4 = 113$. For males, $Q_1 = 69, Q_2 = 80, Q_3 = 87, Q_4 = 112$. A box and whisker plot of this data can be seen in Figure 4. Approximately 48.087% of women are dead before they turn 84 and 49.2% of men are dead by 80.

Figure 4: Probability of Death for Males and Females by Year



The mean age of death for males and females was calculated using a the sum of weighted averages, $\bar{x} = \sum_{x=0}^{119} x * p(x)$. These values where then floored. The sample mean for males was 75 and 36.314% of men die before this age. The sample mean for females is 80. 35.941% of females die before the average age of death.

The standard deviation for each sex was found using

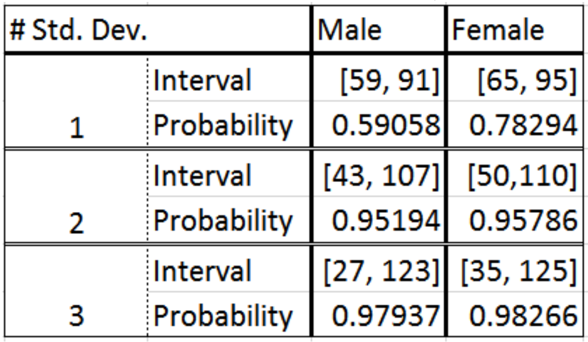
$$s_x = \sqrt{(\sum_{x=0}^{119} x^2 * p(x)) - \mu^2}$$

For women, the variance is $s^2 = 235.514$ and the standard deviation is $s = 15.345466$. For men, the variance is $s^2 = 283.0983$ and the standard deviation is $s = 16.8255$.

Using this information, I found the equivalent of the Empirical Rule for these distributions. For each number of standard deviations from the mean, j, I used Excel to find the cumulative probability of the range $[\bar{x} - \lfloor s \rfloor * j, \bar{x} + \lfloor s \rfloor * j]$.

The resulting values can be seen in Table 5. For men, 59.058% die within one standard deviation of the mean, 95.194% die withing two standard deviations of the mean, and 97.937% die within three standard deviations of the mean. For women, 78.294% die within one standard deviation of the mean, 95.786% die withing two standard deviations of the mean, and 98.266% die within three standard deviations of the mean.

Table 5: Empirical Rule with Mortality Distributions



# Std. Dev.		Male	Female
1	Interval	[59, 91]	[65, 95]
	Probability	0.59058	0.78294
2	Interval	[43, 107]	[50,110]
	Probability	0.95194	0.95786
3	Interval	[27, 123]	[35, 125]
	Probability	0.97937	0.98266

Observing the different maximums in Figure 3 lead me to question whether there was a significant difference between the mean lifespans of males and females. I conducted a test with $H_0 : \mu_{males} = \mu_{females}$, $H_1 : \mu_{males} \neq \mu_{females}$ with significance level $\alpha = 0.050$. The critical region can be calculated using the following equation:

$$|\bar{x} - \bar{y}| \geq t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where

$$S_p = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_Y^2}{n+m-2}} \quad .$$

I inserted the mortality statistics such that x = male age, y = female age, and $t_{\alpha/2} = z_{0.025} = 1.960$, such that

$$S_p = \sqrt{\frac{(100000-1)283.0983226236 + (100000-1)235.51402368}{100000+100000-2}}$$
$$= 16.10298 \quad .$$

Thus the critical region is

$$|75.83 - 80.614| \geq 1.960 * (100000 + 100000 - 2) * 16.10298 * \sqrt{\frac{1}{100000} + \frac{1}{100000}}$$
$$4.784 \ngeq 28229.7793$$

Since $4.784 < 28229.7793$, the null hypothesis is clearly rejected. This is unsurprising given the large sample size of this data set, which would cause any difference to be significant.

The primary analyzers of mortality tables are life insurance companies. I decided to try my hand as an actuary to determine premiums for 20-year, $100,000 policies such that the company breaks even on the policy. To do this, I created a wrote a Python program which optimizes the premium such that the company's profits would average between 0 and 19 cents per customer. While the net policy profit is not in this range, one cent is added or removed from the premium. This premium is then tested in a function that calculates the average payout and revenue collected each year using the cumulative probability of death during the policy to find the net profit. This code can be viewed in the Appendix or at `https://github.com/SandersKM/AnnualPremiums`. The GitHub also contains the probability values for both sexes in a CSV.

I first decided to find my own annual premium. As a 20 year old woman, my likelihood of dying in the next 20 years is rather low. My annual premium should be $70.93. At this rate, the insurance company would make an average of $0.05 per customer.

I then found the annual premium for Dr. Camfield, a 38 year old man. He should pay $430.81 per year. This rate gives the insurer a profit of $0.09 per policy holder.

Lastly, I found the annual premium for Dr. Campbell, a 58 year old man. Due to the risk associated with his policy, his annual premiums are fairly high. He should pay $1,791.24 per year. At this rate, the insurance company would profit $0.10 per customer.

Appendix

```
import random
import locale

def main():
    #csv with mortality rates for ages 0-119. No labels.
    #format: "Male prob, Female prob \n"
    file = "mortalityrates.csv"
    prob_file = open(file, "r")
    prob = prob_file.readlines()
    locale.setlocale(locale.LC_ALL, ’’)
    print("Welcome to the annual premium calculator!")
    print("Determine the annual premium each policy holder should pay
        for your insurance company to break even.")

    finished = False
    while not finished:
        policy, cumdead = get_data(prob)
        premium, net = optimize(policy, cumdead)
        print("The annual premium for this policy holder should be " +
             locale.currency(premium, grouping=True)+ ".")
        print("At this rate, the profit per policy holder would be
            around "+
             locale.currency(net, grouping=True)+ ".")
        done = input("Would you like to look at another policy? ")
        if done[0].lower() == "n":
            print("Thank you for using the annual premium calculator!")
            finished = True

def get_data(prob):
    #Male = 0, Female = 1
    sex = None
    while sex == None:
        sex_input = input("Enter the sex (M/F) of the policy holder: ")
        if sex_input[0].lower() == "m":
            sex = 0
        elif sex_input[0].lower() == "f":
            sex = 1
    age = "NO"
    while (not age.isdigit()):
        age = input("Enter the integer age of the policy holder: ")
    age = int(age)
    policy = []
    cum_dead = [0]
    i = 0
    dead = 0
    while i < 20:
        policy.append(float(prob[age + i].split(",")[sex]))
        dead += float(prob[age + i].split(",")[sex])
        cum_dead.append(dead)
        i += 1
    return (policy, cum_dead)

def optimize(policy, cum_dead):
    premium = (sum(policy) * 100000)/20
    finished = False
    i = 0
    while not finished:
        net = test_premium(premium, policy, cum_dead)
        if net < .19 and net > 0:
            finished = True
        elif net < 0:
            premium += .01
        else:
            premium -= .01
    return (premium, net)

def test_premium(premium, policy, cum_dead):
    net = 0
    for i in range(len(policy)):
        net += (premium * (1-cum_dead[i]))
    net -= cum_dead[-1] * 100000
    return net

main()
```